



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
-----------------	-------------	----------------------	---------------------	------------------

10/605,631

10/15/2003

Alain Franciosa

D/A3358Q

2630

25453

7590

11/02/2006

PATENT DOCUMENTATION CENTER

XEROX CORPORATION

100 CLINTON AVE., SOUTH, XEROX SQUARE, 20TH FLOOR
ROCHESTER, NY 14644

EXAMINER

SAEED, USMAAN

ART UNIT

PAPER NUMBER

2166

DATE MAILED: 11/02/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.		Applicant(s)	
	10/605,631		FRANCIOSA ET AL.	
	Examiner		Art Unit	
	Usmaan Saeed		2166	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

- 1) ☒ Responsive to communication(s) filed on 07 August 2006.
- 2a) ☐ This action is **FINAL**. 2b) ☒ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

- 4) ☒ Claim(s) 1-20 is/are pending in the application.
- 4a) Of the above claim(s) _____ is/are withdrawn from consideration.
- 5) ☐ Claim(s) _____ is/are allowed.
- 6) ☒ Claim(s) 1-20 is/are rejected.
- 7) ☐ Claim(s) _____ is/are objected to.
- 8) ☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

- 9) ☐ The specification is objected to by the Examiner.
- 10) ☒ The drawing(s) filed on 15 October 2003 is/are: a) ☒ accepted or b) ☐ objected to by the Examiner.
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
- a) ☐ All b) ☐ Some * c) ☐ None of:
1. ☐ Certified copies of the priority documents have been received.
2. ☐ Certified copies of the priority documents have been received in Application No. _____.
3. ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

- | | |
|--|---|
| 1) <input checked="" type="checkbox"/> Notice of References Cited (PTO-892) | 4) <input type="checkbox"/> Interview Summary (PTO-413) |
| 2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948) | Paper No(s)/Mail Date. _____ |
| 3) <input checked="" type="checkbox"/> Information Disclosure Statement(s) (PTO/SB/08) | 5) <input type="checkbox"/> Notice of Informal Patent Application |
| Paper No(s)/Mail Date <u>7/3/06</u> . | 6) <input type="checkbox"/> Other: _____ |

DETAILED ACTION

Response to Amendment

1. Applicant's request for reconsideration, filed on 8/07/2006 is acknowledged.
Claim 11 has been amended.

Specification

2. The amended specification was received on 8/07/2006 and is acceptable.

Claim Rejections - 35 USC § 101

3. Regarding claims 11-19, examiner has withdrawn the 101 rejection.

Claim Rejections - 35 USC § 103

4. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negated by the manner in which the invention was made.

Art Unit: 2166

This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

Claims 1-2, 4-5, 10-12, 14-15, and 20 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Steven J. Simske**. (**Simske** hereinafter) (U.S. PG Pub No. 2004/0133560) in view of **Henkin et al.** (**Henkin** hereinafter) (U.S. PG Pub No. 2002/0107735).

With respect to claim 1, **Simske** teaches a method for computing a measure of similarity between a first (or input) document and a second (or search results) document, comprising:

“(a) receiving a first list of rated keywords extracted from the first document and a second list of rated keywords extracted from the second document” as organizing electronic documents may include generating a list of weighted keywords for each document (**Simske** Abstract, & Fig. 4).

Art Unit: 2166

“(b) using the first and second lists of rated keywords to determine whether the first document forms part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list” as the clustering process begins when the weighted keyword lists of two or more documents are compared (step 601). The host device calculates a value, called "shared word weight," that correlates the two documents. The shared word weight value indicates the extent to which two or more documents are related based on their keywords. A higher shared word weight indicates that the documents are more likely to be related (**Simske Paragraph 0048**).

“(c) computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that the first document is included in the second document” as another possible way of weighting the relevancy metrics is to multiply the mean shared weight of extended words shared by two selected text units, e.g., sentences, by the frequency metric of the shared extended words, i.e., the mean ratio of the extended word occurrences in the two documents compared to their occurrences in the larger corpus (**Simske Paragraph 0064**).

“(d) using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed percentage” as clustering documents with common titles, using weighted keywords to determine similarities between documents, etc., a preferred method uses a

Art Unit: 2166

threshold shared word weight and a maximum, mean, or minimum shared word weight as explained above (**Simske** Paragraph 0055).

Simske teaches the elements of claim 1 as noted above but does not explicitly teach, “**percentage of keywords and neighboring keywords.**”

However, **Henkin** discloses “**percentage of keywords and neighboring keywords**” as (**Henkin** Paragraph 0222 & 0288).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Henkin’s** teachings would have allowed **Simske** to determine the minimum percentage of matched words to be found in the document context in order to conclude that a match exists.

Claims 11 and 20 are essentially the same as claim 1 except they set forth the claimed invention as a system and an article of manufacture and are rejected for the same reasons as applied hereinabove.

With respect to claim 2, **Simske** teaches “**the method according to claim 1, wherein the second percentage at (c) is computed by giving weight only to those keywords and their set of neighboring keywords in the first list that match in the second list and a threshold percentage of the keywords in their set of neighboring keywords**” as shown in Table 5, the documents share two keywords,

Art Unit: 2166

"Hockey" and "Skating." The shared word weight value of the keywords may be chosen in a variety of ways, e.g., maximum, mean, and minimum (**Simske** Paragraph 0050).

Claim 12 is essentially the same as claim 2 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

With respect to claim 4, **Simske** teaches **"the method according to claim 2, wherein the threshold percentage is reduced when the first list of rated keywords is identified using OCR"** as the documents included in each cluster may be adjusted by changing the threshold of the required shared word weight for clustering (**Simske** Paragraph 0058). If any documents being considered are paper-based, tools such as a zoning analysis engine in combination with an optical character recognition (OCR) engine may be used to convert the paper-based document to an electronic document (**Simske** Paragraph 0016).

Claim 14 is essentially the same as claim 4 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

With respect to claim 5, **Simske** teaches **"the method according to claim 1, further comprising (e) if the first computed percentage does not indicate that the first document is included in the second document, computing a third percentage using the Jaccard distance measure"** as the shared word weight value indicates the

Art Unit: 2166

extent to which two or more documents are related based on their keywords. A higher shared word weight indicates that the documents are more likely to be related (**Simske** Paragraph 0048). Examiner interprets that when a document is related/included in a second document it does not need to calculate third percentage.

Alternatively if it does not indicate then examiner interprets the above limitation as the relevance weight for A is calculated, as shown, by summing (step 704), the weight of B divided by the distance of B (as measured in characters) from A (step 703), the weight of C divided by the distance of C from A (step 703), the weight of D divided by the distance of D from A (step 703), then multiplying that sum by the weight of A (step 705). The summation of keyword weights divided by their respective distances to a particular occurrence can be called a "distance metric" (step 704) (**Simske** Paragraph 0062).

Claim 15 is essentially the same as claim 5 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

With respect to claim 10, **Simske** teaches, **"the method according to claim 1, wherein the first document is a portion of the second document"** as a method and system for organizing electronic documents by generating a list of weighted keywords, clustering documents sharing one or more keywords, and linking documents within a cluster by using similar keywords, sentences, paragraphs, etc., as links. The embodiments provide customizable user control of keyword quantities, cluster

Art Unit: 2166

selectivity, and link specificity, i.e., links may connect similar paragraphs, sentences, individual words, etc (**Simske** Paragraph 0015).

5. Claims 3, 6-7, 13, and 16-17 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Steven J. Simske**. (U.S. PG Pub No. 2004/0133560) and **Henkin et al.** (U.S. PG Pub No. 2002/0107735) as applied to claims 1-2, 4-5, 10-12, 14-15, and 20 above, further in view of **Rie Kubota**. (**Kubota** hereinafter) (U.S. Patent No. 6,041323).

With respect to claim 3, **Simske** teaches “**the method according to claim 2, wherein the second percentage at (c) is computed by giving full weight to those keywords in the first list of rated keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords**” as the experiment consists of varying the weighting, e.g., ranging the weight from 0.1 to 10.0 using 0.1 steps, for a particular attribute (**Simske** Paragraph 0032). Examiner considers 10 as being full weight.

Simske teaches the elements of claim 3 as noted above but does not explicitly disclose “**keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords.**”

However, **Kubota** discloses “**keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords**” as the fixed length chain is searched from the character chain file. In step 508, if it is determined that no fixed length chain is found, a message box is preferably

Art Unit: 2166

displayed in step 526 for indicating that the search character string cannot be found, and the process ends (**Kubota** Col 26, Lines 44-48). Therefore the reference teaches that keywords are not found in the second set of keywords/document.

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Kubota's** teachings would have allowed **Simske and Henkin** to provide a search method, which requires less storage capacity and extracts a unique character string at a high speed (**Kubota** Col 2, Lines 51-53) and to provide a method for searching for a comparison document, which has character strings similar to a partial input character string existing in an input document (**Kubota** Col 2, Lines 3-6).

Claim 13 is essentially the same as claim 3 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

With respect to claim 6, **Simske** does not explicitly teaches "**the method according to claim 5, further comprising (f) if the third computed percentage indicates that the first document is a revision of the second document, computing a fourth percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the second list also exist in the first list.**"

However, **Kubota** discloses "**the method according to claim 5, further comprising (f) if the third computed percentage indicates that the first document**

Art Unit: 2166

is a revision of the second document, computing a fourth percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the second list also exist in the first list” as in the case of multiple documents, it may be a set of documents including the input document, or a set of documents extracted by search or the like (**Kubota** Col 3, Lines 63-66). Examiner interprets the input document as revision since the input document is the output to the search using keywords in the input document.

Calculating the similarity factor of the comparison document from the first appearance frequency value taking the first weight value into account and the second appearance frequency value taking the second weight value into account (**Kubota** Col 6, Lines 22-25). Examiner interprets comparison document as second document.

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Kubota's** teachings would have allowed **Simske and Henkin** to provide a search method, which requires less storage capacity and extracts a unique character string at a high speed (**Kubota** Col 2, Lines 51-53) and to provide a method for searching for a comparison document, which has character strings similar to a partial input character string existing in an input document (**Kubota** Col 2, Lines 3-6).

Claim 16 is essentially the same as claim 6 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

With respect to claim 7, **Simske** teaches “**the method according to claim 6, further comprising using the fourth computed percentage to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords**” as if any documents being considered are paper-based, tools such as a zoning analysis engine in combination with an optical character recognition (OCR) engine may be used to convert the paper-based document to an electronic document (**Simske** Paragraph 0016). The keywords in the documents are being identified using OCR in the reference. Therefore, there is no need for using fourth computed percentage to specify the measure of similarity.

Claim 17 is essentially the same as claim 7 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

6. Claims 9 and 19 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Steven J. Simske**. (U.S. PG Pub No. 2004/0133560) and **Henkin et al.** (U.S. PG Pub No. 2002/0107735) as applied to claims 1-2, 4-5, 10-12, 14-15, and 20 above, in view of **Drissi et al.** (**Drissi** hereinafter) (U.S. PG Pub No. 20003/0149686).

With respect to claim 9, **Simske** does not explicitly teaches “**the method according to claim 1, wherein the first list of rated keywords includes one or more keywords translated from a second language different from a first language that is identified as being a primary language of the first document.**”

However, **Drissi** discloses “**the method according to claim 1, wherein the first list of rated keywords includes one or more keywords translated from a second language different from a first language that is identified as being a primary language of the first document**” as an inverted index 214 is created from the translated keywords. The translation of keywords is preferably accomplished using a keyword dictionary 220 which included words in English associated with the corresponding keywords in the national language (and vice versa) to form a synonym listing which effectively translates a keyword in one language into the corresponding term in another language and vice versa) (**Drissi** Paragraph 0024). Examiner interprets the national language as primary language.

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Drissi's** teachings would have allowed **Simske and Henkin** to provide translation process to allow searching of the documents in different languages (**Drissi** Paragraph 0012).

Claim 19 is essentially the same as claim 9 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

Art Unit: 2166

7. Claims 8 and 18 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Steven J. Simske**. (U.S. PG Pub No. 2004/0133560).

With respect to claim 8 **Simske** teaches “**the method according to claim 1, wherein the first computed percentage indicates that the first document is included in the second document when the percentage defined by ratio of $\text{Sum1}/\text{Sum2}$ is greater than approximately ninety percent, where**” as for example, if a threshold shared word weight value of 0.7 is designated, and the two documents of Table 5 are being compared for possible clustering, using the maximum shared word weight value (1.0) will cluster the two documents, while using the mean shared word weight (0.5) or minimum shared word weight values (0.3) will not cluster the two documents (**Simske** Paragraph 0052). Examiner interprets the threshold value of 70 percent as 90 percent.

“**D1 is the number of keywords in first list of keywords**” as table 5 with keywords from document 1 and document 2 (**Simske** Paragraph 0049).

“**D2 is the number of keywords in the second list of keywords**” as table 5 with keywords from document 1 and document 2 (**Simske** Paragraph 0049).

“**Sum1 is the sum of the weights of keywords that appear in D1 that also appear in D2**” as the sum of all weight values for "Hockey" and "Skating" is $0.4+0.25+0.3+0.05=1.0$ (**Simske** Paragraph 0052). Hokey and Skating appear in both D1 and D2.

“Sum2 is the sum of the weights of keywords in D1” as the keywords are located, a sentence weight is calculated (502), for example, by adding together all the keyword weights (**Simske** Paragraph 0045).

Simske teaches the elements of claim 8 as noted above but does not explicitly disclose **“Sum1/Sum2.”**

However, **Simske** teaches **“Sum1/Sum2”** as the mean shared word weight value is $[\text{fraction } (1.0/2)]=0.5$ (**Simske** Paragraph 0052).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of the cited reference to find the ratio for two possible similar documents by dividing the sum of keywords from both documents by sum of keywords in one document.

Claim 18 is essentially the same as claim 8 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

Response to Arguments

8. Applicant's arguments have been considered but are moot in view of the new ground(s) of rejection.

Applicant argues that **Simske** does not teach or suggest **“using the first and second lists of rated keywords to determine whether the first document forms**

Art Unit: 2166

part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list,” “computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that the first document is included in the second document” and “using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed percentage.”

In response to applicant arguments examiner respectfully submits that **Simske** teaches **“(b) using the first and second lists of rated keywords to determine whether the first document forms part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list”** as the clustering process begins when the weighted keyword lists of two or more documents are compared (step 601). The host device calculates a value, called "shared word weight," that correlates the two documents. The shared word weight value indicates the extent to which two or more documents are related based on their keywords. A higher shared word weight indicates that the documents are more likely to be related (**Simske** Paragraph 0048).

“(c) computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the first computed percentage indicates that

Art Unit: 2166

the first document is included in the second document” as another possible way of weighting the relevancy metrics is to multiply the mean shared weight of extended words shared by two selected text units, e.g., sentences, by the frequency metric of the shared extended words, i.e., the mean ratio of the extended word occurrences in the two documents compared to their occurrences in the larger corpus (**Simske Paragraph 0064**).

“(d) using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed percentage” as clustering documents with common titles, using weighted keywords to determine similarities between documents, etc., a preferred method uses a threshold shared word weight and a maximum, mean, or minimum shared word weight as explained above (**Simske Paragraph 0055**).

Simske teaches the elements of claim 1 as noted above but does not explicitly teach, **“percentage of keywords and neighboring keywords.”**

However, **Henkin** discloses **“percentage of keywords and neighboring keywords”** as (**Henkin Paragraph 0222 & 0288**).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Henkin’s** teachings would have allowed **Simske** to determine the minimum percentage of matched words to be found in the document context in order to conclude that a match exists.

Art Unit: 2166


Contact Information

9. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Usmaan Saeed whose telephone number is (571)272-4046. The examiner can normally be reached on M-F 8-5.

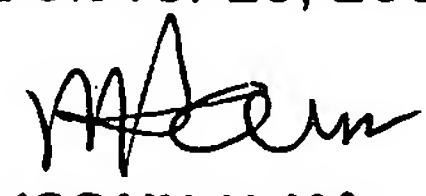
If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Hosain Alam can be reached on (571)272-3978. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Usmaan Saeed
Patent Examiner
Art Unit: 2166

Leslie Wong 
Primary Examiner

US
October 26, 2006


HOSAIN ALAM
SUPERVISORY PATENT EXAMINER